2014

Detection of temporal lobe epilepsy using support vector machines in multi-parametric quantitative MR imaging.

Diego Cantor-Rivera

Ali R Khan

Maged Goubran

Seyed M Mirsattari

Terry M Peters

Detection of Temporal Lobe Epilepsy using Support Vector Machines in Multi-parametric Quantitative MR Imaging

Diego Cantor-Rivera[a,c,*], Ali R. Khan[a,b], Maged Goubran[a,c], Seyed M. Mirsattari[d], Terry M. Peters[a,b,c]

[a] *Imaging Research Laboratories, Robarts Research Institute, London, ON, Canada*
[b] *Dept. of Medical Biophysics, Western University, London, ON, Canada*
[c] *Biomedical Engineering Graduate Program, Western University, London, ON, Canada*
[d] *Dept. of Clinical Neurological Sciences, Medical Biophysics, Medical Imaging and Psychology, Western University, London, ON, Canada*

Abstract

The detection of MRI abnormalities that can be associated to seizures in the study of temporal lobe epilepsy (TLE) is a challenging task. In many cases, patients with a record of epileptic activity do not present any discernible MRI findings. In this domain, we propose a method that combines quantitative relaxometry and diffusion tensor imaging (DTI) with support vector machines (SVM) aiming to improve TLE detection. The main contribution of this work is two-fold: on one hand, the feature selection process, principal component analysis (PCA) transformations of the feature space, and SVM parameterization are analyzed as factors constituting a *classification model* and influencing its quality. On the other hand, several of these classification models are studied to determine the optimal strategy for the identification of TLE patients using data collected from multi-parametric quantitative MRI. R3

A total of 17 TLE patients and 19 control volunteers were analyzed. Four images were considered for each subject (T1 map, T2 map, fractional anisotropy, and mean diffusivity) generating 936 regions of interest per subject, then 8 different classification models were studied, each one comprised by a distinct set of factors. Subjects were correctly classified with an accuracy of 88.9%. Further analysis revealed that the heterogeneous nature of the disease impeded an optimal outcome. After dividing patients into cohesive groups (9 left-sided seizure onset, 8 right-sided seizure onset) perfect classification for the left group was achieved (100% accuracy) whereas the accuracy for the right group remained the same (88.9%).

We conclude that a linear SVM combined with an ANOVA-based feature selection + PCA method is a good alternative in scenarios like ours where feature spaces are high dimensional, and the sample size is limited. The good accuracy results and the localization of the respective features in the temporal lobe suggest that a multi-parametric quantitative MRI, ROI-based, SVM classification could be used for the identification of TLE patients. This method has the potential to improve the diagnostic assessment, especially for patients who do not have any obvious lesions in standard radiological examinations.

Keywords: MRI, DESPOT, DTI, FA, MD, quantitative imaging, feature selection, mRMR, ANOVA, support vector machines, machine learning, SVM, ROI, PCA, TLE, epilepsy

1. Introduction

Magnetic resonance imaging (MRI) is a powerful tool for the evaluation of patients with brain disorders and neurological diseases. For those with temporal lobe epilepsy (TLE), the most common type of epilepsy in adults[1], it is the entry point to a clinical workflow that may conclude with temporal lobe surgery and an improved quality of life. Finding evidence of seizures on MRI is a clear diagnostic element for TLE, however this task is not easy given that epileptogenic lesions are often small and can be missed, they can be uncertain due to subtle intensity changes, or only perceptible after image post-processing. Furthermore, due to the multi-factorial nature of the disease, the localization and type of the lesions can vary from patient to patient.

While the visual inspection is a common radiological procedure for the diagnosis of TLE, it has been shown that the detection of brain pathologies associated with TLE can be improved with computer-assisted, automatic multi-parametric MRI analysis. For example, the detection of changes in the shape, volume and intensity of the hippocampus has been studied using structural T1- and T2- weighted images [2, 3, 4, 5]; White matter abnormalities have been detected with diffusion tensor imaging (DTI) in TLE patients [6, 7, 8], and DTI has been employed concurrently with functional MRI (fMRI) to perform language lateralization of TLE patients [9].

A support vector machine (SVM) is a classifier that uses *a priori* knowledge in the form of group labels (supervised learning) and produces a *decision boundary* that can be used to determine the label of new examples [10, 11]. Recent studies have examined the possibility of improving TLE detection using SVMs on MRI data. For example Focke et al. [12] show correct patient lateralization (left vs. right seizure onset) using SVMs on T1-weighted and DTI data. In addition to lateralization, Keihaninejad et al. [13], demonstrate the identification of TLE cases with hippocampal atrophy from cases without it using SVM on regional volumes obtained from T1-weighted MRI.

In this context, the goal of the current study is to explore TLE detection using multi-parametric quantitative MRI and support vector machines. For that purpose, quantitative maps of T1, T2, as well as fractional anisotropy (FA), and mean diffusivity (MD) are estimated for every participating subject. Quantitative MRI measures biophysical tissue properties and it has the potential to be more sensitive to TLE detection than T1- and T2-weighted images. Additionally, quantitative measurements are independent of experimental settings and thus comparable between different scanners, institutions and over different points in time [14].

*Corresponding author
Email address: dcantor@robarts.ca (Diego Cantor-Rivera)

2. Methods

2.1. Overview

All subjects in this study underwent an imaging protocol approved by the Office of Research Ethics of Western University (Canada). The imaging protocol comprised DESPOT1, DESPOT2 [15, 16] and DTI sequences, resulting image data being processed to obtain four different quantitative maps: T1, T2, fractional anisotropy (FA) and mean diffusivity (MD). Anatomical atlas-based labeling was used to define ROIs on each one of the quantitative maps and subsequently to measure and extract regional features.

Given that the number of features exceeded the number of subjects (936 features, 36 subjects), two different feature selection methods were explored to discard irrelevant or uninformative features. Then, the feature space was further reduced using principal component analysis (PCA).

An SVM was trained/tested on the filtered feature space using a leave-one-out cross-validation strategy (LOOCV), where the SVM was tasked with predicting the label (patient or control) for the omitted subject (Figure 1). Once all subjects were evaluated, sensitivity, specificity and classification accuracy were measured and reported. This procedure was repeated for each *classification model* obtained by the combination of the following elements:

- the image that originates the features (T1, T2, FA, MD, or all combined)
- the method to select features
- the cardinality of the requested feature set [K]
- the use of PCA to reduce the feature space
- the type of SVM

A detailed comparison among classification models is reported in the results section along the best classification scenarios. Specific recommendations regarding the optimal model are given in the discussion section. In addition, an analysis of the elements constituting a classification model and their influence in the classifiers performance is also discussed. Finally, the features relevant for classification are analyzed and their clinical significance is considered.

2.2. Participants

Thirty-six individuals participated in this study, 19 of whom were control volunteers (age 32±10,12 male, 7 female) and 17 TLE patients (age 35±10, 8 male, 9 female). All the patients had lateralizable seizures (confirmed by EEG) and all of them were eligible for temporal lobectomy (9 left, 8 right). Preoperative MRI and post-surgical pathology confirmed the presence of Mesial Temporal Sclerosis (MTS) in 8 of them.

2.3. Imaging Protocol and Preprocessing

Subjects were scanned (presurgically in the case of patients) using a 3T MR scanner (GE Discovery MR750) with whole brain, 1mm isotropic DESPOT1-HiFi and DESPOT2-FM ,T1 and T2 mapping sequences [15, 16] respectively, optimized for imaging at 3T [17]. Two SPGR images were acquired (flip angles of 4° and 18°) along with an inversion-prepared spoiled gradient recalled acquisition in the steady state (SPGR) to calculate a quantitative T1 map. Five balanced steady state free-precession (bSSFP) images were acquired with phase cycling (flip angles of 15°,35°,60°) to estimate a quantitative T2 map using the DESPOT2-FM procedure [18]. All images were co-registered to the 18° flip angle SPGR image using the FLIRT registration tool [19] from the FSL software (`http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/`) to account for motion between scans prior to the computation of the T1 and T2 maps.

A DTI sequence was also acquired with the following parameters: 2.4mm isotropic, 41 directions, b-value =1000, 4 non-weighted (b=0) volumes. Non-linear distortions were corrected by deformable registration of the average of the b=0 volumes to the undistorted T1 map, using a diffeomorphic registration method [20, 21]. Eddy current correction and diffusion tensor estimation was performed using FSL's diffusion toolbox FDT. Maps of fractional anisotropy (FA) and mean diffusivity (MD) were transformed and re-sampled to the coordinate system defined by the 1mm isotropic T1 map. Synthetic T1-weighted images, with inherent bias-field correction, were generated from the T1 maps [22] and used in place of directly acquired T1-weighted images for subsequent segmentation.

This preprocessing stage yielded four co-registered quantitative maps (T1, T2, FA and MD) for each subject.

2.4. Feature Extraction

Volumetric segmentation of the synthetic T1-weighted images was performed with Freesurfer (`http://surfer.nmr.mgh.harvard.edu/`). This processing included removal of non-brain tissue using a hybrid watershed/surface deformation procedure, automated Talairach transformation, segmentation of the subcortical white matter and deep gray matter volumetric structures (including hippocampus, amygdala, caudate, putamen, ventricles). Once the cortical models were completed, the cerebral cortex was parcellated into regions based on gyral and sulcal structures [23]. This segmentation produced bilateral regions of interest for every subject in subject image space. Total of 35 cortical, 34 white matter and 9 subcortical regions (78 in all) were identified per hemisphere. The segmented ROIs were used to extract features from the quantitative maps (T1, T2, FA, and MD) for each subject (Figure 3).

Two types of features were extracted: The ROI *mean intensity*, and the *asymmetry* between correspondent left and right ROIs (i.e. left and right hippocampus). The asymmetry was expressed as the non-parametric two-sample Kolmogorov-Smirnov statistical score between left-right region pairs. For each subject, a vector of 936 features was created, accounting for 624 mean intensity features (156 ROIs, 4 image

sources) and 312 intensity difference features (78 left-right region pairs, 4 image sources). Each feature vector was labelled according to the respective subject ground truth (i.e. patient or control). Prior to using these feature vectors as input for the SVM, irrelevant features were discarded.

2.5. Support Vector Machines

Linear SVMs are used in scenarios where the features that discriminate between subject groups are linearly separable. Linear SVMs estimate the hyperplane or *decision boundary* that provides the largest margin (separation) between the two groups [11]. In contrast, non-linear SVMs such as *radial basis function* SVMs provide a non-linear boundary using the *kernel trick* as described in [10] which effectively transforms the non-linear space into a space that has a higher number of dimensions but where the features are linearly separable. The *scikit-learn* library [24] was used to implement the SVM classifiers in this work. Both linear and non-linear support vector machines were used for patient detection (Figure 2).

2.6. Experiment Set-Up

Each subject was assigned one of two possible labels: *patient* or *control*. The SVM was trained using feature vectors to distinguish between these two groups. A leave-one-out cross-validation approach (LOOCV) was used for training/testing. One subject was left out from the training stage while the remainder were used for training (i.e. to compute the decision boundary). Then, the classifier was asked to predict the correct label for the excluded subject. This procedure iterated until all subjects had been classified. Afterwards, the *sensitivity*, *specificity* and *accuracy* were calculated as follows: R4

$$sensitivity = \frac{tp}{tp + fc} \tag{1}$$

$$specificity = \frac{tc}{fp + tc} \tag{2}$$

$$accuracy = \frac{tp + tc}{p + c} \tag{3}$$

where tp = correctly identified patients, tc = correctly identified controls, fp = controls misclassified as patients, fc = patients misclassified as controls, p = total number of patients, c = total number of controls.

As shown in the results section, patients were subsequently split into those with left temporal lobe seizures and those with right temporal lobe seizures (L-TLE and R-TLE respectively). The same training/testing procedure was repeated to assess the detectability of each patient subgroup.

2.7. Classification Models

The outcome of an SVM is influenced by the feature space over which it operates (extrinsic element) as well as by the type of SVM (intrinsic element). In this work, we define a *classification model* as the combination of *extrinsic* and *intrinsic* elements selected for the operation of the SVM.

Each model is built in three steps as shown in Figure 1: One of two possible feature selection methods (correlation-based or ANOVA-based); the option or not of performing dimensionality reduction with PCA; and the type of SVM (linear or non-linear). Eight different classification models were evaluated:

- *correlation-svm-linear:* Features are selected with a correlation-based method and a linear SVM runs on this space.
- *correlation-pca-svm-linear*: After the correlation-based feature selection, the dimension of the resulting feature space is reduced using PCA. A linear SVM is used for classification.
- *correlation-svm-rbf*: Features are selected with a correlation-based method and the SVM used is non-linear (radial basis functions).
- *correlation-pca-svm-rbf*: The dimension of the feature space obtained with correlation-based selection is simplified using PCA. The employed SVM is non-linear.
- *anova-svm-linear*: Features are selected using an ANOVA-based approach. The SVM used is linear.
- *anova-pca-svm-linear*: Features are selected using an ANOVA-based approach. The resulting feature space is transformed using PCA. The SVM used is linear.
- *anova-svm-rbf*: Features are selected using an ANOVA-based approach. The SVM is non-linear.
- *anova-pca-svm-rbf*: Similar to the previous case, but the resulting feature space is reduced with PCA.

The following sections describe the steps taken to build a classification model, while the results section compares models based on their performance. Finally, the best and worst models considering feature selection, PCA, cardinality and computational performance elements are discussed and an analysis of relevant, stable features and their clinical significance is presented.

2.8. Feature Selection

Feature selection algorithms identify features that are pivotal for classification [25], and optimize the accuracy of machine learning algorithms (such as SVMs), while avoiding overfitting of the data [26].

In theory, feature selection should be performed on data that are independent of both the training and testing sets [27, p. 222]. In practice, this is not always feasible, particularly when the number of features is much larger than the number of subjects, as is the case in our work. To address this issue, our feature

selection step subsamples the training set, creating leave-one-subject-out folds and then scores features on each resulting fold. After this, a voting algorithm examines the agreement among the folds, and decides which features belong to the final feature set. This approach reuses the data efficiently and reduces the outlier effect (performance reduction attributable to a subject).

2.8.1. Training Set Subsampling

The training set was subsampled by leaving one subject out every time, creating M folds for a training set of size M (Figure 4). On each fold, one of two possible feature selection methods (correlation or ANOVA-based) was evaluated. Each method began by assigning a score to each of the 936 features for each subject in the current fold. Then, using the *cardinality parameter* K, set by the user, the respective method selected the top $2K$ features based on their scores. Finally, after all folds were evaluated, the voting algorithm decided on the final K features by reviewing the agreement among the folds.

2.8.2. Correlation-Based Feature Selection

Correlation-based feature selection analyzes the linear dependency between features and classification groups. Features that have a high correlation with the group labels (relevance) are good candidates for the prediction of the subject's group [25]. Similarly to previous work, our approach accounts for inter-feature correlation. Features with high inter-feature correlation (redundancy) are penalized producing a set where features tend to be linearly independent of each other [28, 29, 30].

The correlation-based feature selection, approached in three steps: relevance evaluation, redundancy evaluation, and final scoring, is applied to each subsampled training fold.

First, *feature relevance* is evaluated using a Pearson correlation coefficient. Let N $(= M - 1)$ be the total number of subjects in the current fold; $f_{i,s}$ the value of the feature i evaluated in the subject s; and t_s the binary group label (-1,+1) for subject s (for binary classification). Then, the relevance for feature i is obtained by:

$$\rho(i) = \frac{\sum_{s=1}^{N}(f_{i,s} - \bar{f}_i)(t_s - \bar{t})}{\sqrt{\sum_{s=1}^{N}(f_{i,s} - \bar{f}_i)^2 \sum_{s=1}^{N}(t_s - \bar{t})^2}} \tag{4}$$

where $\bar{f}_i$ and $\bar{t}$ are the feature and the label averages respectively, evaluated using all the subjects in the current fold. After this, the top $2K$ features are selected.

In the second step, *feature redundancy* is calculated as the average of the correlation between each feature and the remaining $2K - 1$ features selected on the first step (inter-feature correlation). Features that have a high redundancy do no provide additional information due to high collinearity, and can be discarded. Let

i and j be two different features ($i \neq j$), the inter-feature correlation, in the current fold, is calculated as it follows:

$$\delta(i,j) = \frac{\sum_{s=1}^{N}(f_{i,s} - \bar{f}_i)(f_{j,s} - \bar{f}_j)}{\sqrt{\sum_{s=1}^{N}(f_{i,s} - \bar{f}_i)^2 \sum_{s=1}^{N}(f_{j,s} - \bar{f}_j)^2}} \tag{5}$$

Then, the redundancy for feature i in the current fold, is calculated as:

$$\delta(i) = \frac{\sum_{j=1;j\neq i}^{2K}\delta(i,j)}{2K-1} \tag{6}$$

The third step computes the score, where for each feature in the $2K$ set, the feature-label correlation (relevance) is divided by the respective average inter-feature correlation (redundancy). The scoring function S for feature i in the current fold, was then defined as:

$$S(i) = \frac{\rho(i)}{\delta(i)} \tag{7}$$

Finally, the $2K$ scored features are selected as the feature set for the current fold.

2.8.3. ANOVA-Based Feature Selection

To the best of our knowledge, there is very little literature on the use of ANOVA as a feature selection method for classification of structural brain images. The use of ANOVA for feature selection in neuroimaging, has mainly focused on classification of fMRI datasets [31, 32, 33]. Nonetheless, ANOVA has be used in others areas of science to attenuate the *curse-of-dimensionality* [26] by discarding variables with poor statistical significance thereby reducing the size of the feature space. We explored this approach as an alternative to the more commonly used correlation-based method. Similar to the latter, the ANOVA approach is supervised feature selection method where the *a priori* knowledge of the subject groups (ground truth) is used in the evaluation of the features. Unlike the correlation-based method introduced earlier, the ANOVA-based method does not penalize redundant features.

This method proceeds as follows: The subjects in the current training set fold (as defined above) are first divided into groups according to the classification labels of the given experiment (i.e. patients vs. controls). Then, a one-way ANOVA test is performed for each feature between the groups. The *null hypothesis* for the ANOVA test is that the mean value for each feature is the same between the groups. If there is evidence that a feature mean is significantly different between groups, then that feature becomes a good candidate for the prediction of the group. Finally, a scoring function is assigned to quantify the *relevance* of each feature according to the result of its ANOVA test.

For any feature i, with $0 < i \leq T$ where T is 936, the total number of features, ($K < T$), the ANOVA-based score S for feature i in the current fold is defined as:

$$S(i) = 1 - \alpha_i \tag{8}$$

where α_i is the p-value resulting of the correspondent F-test for feature i. In other words, significant features have low *p-value* and a correspondingly high score. Features with the top $2K$ scores are selected to produce the feature set for the current fold.

2.8.4. Voting Strategy

For both the correlation and ANOVA-based methods, the final feature set is built on a feature-by-feature basis examining the scores obtained by each feature among the participating folds. For any given feature i, the *vote* collected from the fold q is given by:

$$V(i,q) = \begin{cases} S(i) & when\ i \in q \\ 0 & when\ i \notin q \end{cases} \tag{9}$$

Thus the *vote* is 0 if the feature i was not scored in fold q, and is $S(i)$, the scoring function, otherwise. The agreement is reached by:

$$A(i) = \frac{\sum_{q=1}^{M} V(i,q)}{M} \tag{10}$$

The *agreement function* $A(i)$ becomes a score average when feature i has been scored across all folds. Otherwise, infrequent features are penalized even if they have scored highly in a given particular fold. The purpose of this penalization is to enforce feature stability, producing feature sets that are robust.

Given that each voting fold scores $2K$ features, after the agreement function has been applied, only features in the top K agreement values are retained. These features become the final feature set that is used for classification.

2.9. Dimensionality reduction

Dimensionality reduction of high-dimensional feature spaces has been used in neuroimaging to reduce computational complexity, by projecting the high dimensional data onto a space of smaller dimensionality without loss of information [34]. Techniques such as principal component analysis (PCA) [35] have been applied directly to unfiltered highly-dimensional feature spaces in classification problems. However it has been shown that the resulting condensed space may carry over noise from original features, negatively affecting classification performance [36].

As an alternative, we present a two-step approach, where the feature space is filtered using the proposed feature selection stage, followed by the application of PCA to the resulting space. Applying PCA to the feature space creates a new space where each dimension is *uncorrelated*. This could reduce noise and improve classification performance on this space [37].

There is, however, a trade-off when using PCA. On one hand, PCA is an *unsupervised* method. This means that class labels are not required. On the other hand, the number of dimensions or *principal components* to which the feature space is reduced to, is a parameter that must be set by the user. To take advantage of the PCA technique and tackle the selection of the number of principal components, this parameter was included in the optimization step described below.

2.10. SVM selection and parameter optimization

A linear SVM has only one parameter C which determines the level of allowed regularization computing the decision boundary. A low C creates a decision boundary with a large margin between the two classes, while a high C attempts to classify all the data points correctly producing a narrow margin. In contrast, a non-linear, radial basis function SVM has two parameters: C, the regularization parameter and γ which determines the weight of individual observations in the resulting decision boundary. A low γ produces higher weighting while a high γ produces lower weighting (Figure 5).

Clearly, the choice of SVM brings along parameters that need to be optimized. Parameter selection can either boost or hinder the generalization ability of the SVM. To determine the optimal parameters, a cross-validated *grid search* was employed on the parameter space as follows: The training set was divided into evaluation and validation partitions using cross-validation (CV) and an SVM was trained using parameters taken from the parameter grid onto the evaluation partition. The validation partition was then used to measure the accuracy of the classifier. This process was repeated for all possible training and validation partitions given by CV to obtain an average measurement of classification accuracy. This CV process was repeated for all possible parameter sets in the parameter grid. Finally the parameter set that produced the best average accuracy was selected as the optimal choice to configure the SVM (Figure 6).

If the model being evaluated included the PCA reduction step, a range for the number of principal components was included into the parameter grid. In that case, parameter optimization occurred on the PCA-reduced space.

3. Results

Each of the eight classification models was evaluated on the full range of possible K values (from 2 to 936 features for the universal feature space (T1+T2+FA+MD), and from 2 to 234 for individual image subspaces). Given that K is the only parameter that needs to be set by the user, an adequate range for the choice of K is addressed in the discussion section.

3.1. Detection of TLE patients

A maximum accuracy of 88.9% for automatic classification between patients and controls was obtained by an *anova-pca-svm-linear* model using 10 features from the T1 image only as shown in Table 1. This model misclassified 3 patients, all of whom were diagnosed with R-TLE, and one control volunteer. In general the classification models that employed features from the T1 image, obtained 81% average accuracy, followed by the MD models with an average accuracy of 75%, T2 models with an average accuracy of 74% and in last place FA models with an accuracy of 67%.

Figures 7 and 8 (I) show regions selected by the best classification model that were common to all subjects. Figure 7 shows left/right mean intensity features while Figure 8 shows regions selected based on asymmetry. Thus, the middle-temporal cortex, the right parahippocampal white matter and the left entorhinal white matter are selected based on their mean intensity. In contrast, the inferior-temporal, middle-temporal and parahippocampal white matter regions are selected based on their asymmetry. These findings are consistent with previous studies where changes in temporal white matter regions have been associated to TLE [6, 8]. Although feature selection is not restricted to the temporal lobe, it is relevant to notice that all the regions shown in the overlay are temporal lobe regions.

We hypothesized that the heterogeneity of the patient group (L-TLE, R-TLE) was a relevant factor that influenced the accuracy of the patient identification process. Its effect can be seen by performing a dimensionality reduction transformation on the feature space (Figure 9). While the control group forms a cluster, the patient groups are sparse. This configuration was consistently reproducible for feature spaces of different sizes and for different data projection techniques including PCA, Isomaps and Multidimensional Scaling. These projections demonstrated that the L-TLE group tends to be linearly separable from the control group, while the R-TLE group intersected with the control group. This configuration could explain why perfect classification is achievable between controls and L-TLE, while classification between controls and R-TLE appeared to be more challenging. To address the heterogeneity matter, classification accuracy was examined using more homogeneous patient subgroups: L-TLE vs. controls (experiment II) and R-TLE vs. controls (experiment III).

3.2. Detection of L-TLE patients

Perfect classification (100% accuracy) was attainable for L-TLE patients (Table 2). This result was reported by the *correlation-pca-svm-rbf* model with MD features only (K=7).However, the *correlation-pca-svm-linear* and *anova-pca-svm-linear* models also obtained good classification accuracy (96.4%) using MD features with low feature set cardinalities. In general, linear models were not outperformed by non-linear models. Overall, PCA improved classification accuracy.

When looking at individual feature subspaces (columns on Table 2), the best accuracies across models were obtained by the MD and T1 subspaces where the average accuracy was 90% for T1, and 94% for MD.

The average classification accuracy across models dropped in the FA and T2 subspaces to 84% and 73% respectively. In general, classifiers had an excellent performance on the universal feature space where the average accuracy was 90%.

Figures 7 and 8 (II) show the regions selected by the best classification method that were common to all subjects. In general, the asymmetry between left and right temporal lobe white matter regions, and the mean intensity of left temporal cortical regions are determined to be relevant. Specifically, the left hippocampus, the left middle-temporal cortex and the left entorhinal white matter regions are selected based on the *mean intensity*, whereas the enthorhinal cortex, the superiotemporal white matter and the temporal pole are selected based on their *asymmetry*. All of these regions clinically correlate to pathologies in L-TLE patients such as mesial temporal sclerosis (MTS) and focal cortical dysplasia (FCD).

3.3. Detection of R-TLE patients

As presumed by the PCA projection, the classification between R-TLE patients and controls was a difficult problem. This was reflected in classification accuracy average of 74% across the different models as well as the high cardinality reported by each model (Table 3).

A possible explanation for these results is that the feature space does not have discriminative features to distinguish the R-TLE patients from controls (evidenced by the PCA projection). Hence, reliable decision boundaries cannot be obtained. There is some evidence supporting the proximity of R-TLE patients and controls in this type of images. For example, Zhong Xue et al. [6] have found fewer regions to distinguish R-TLE patients from controls than L-TLE from controls. Similarly, Ahmadi et al. [7] have reported fewer and less extensive gray matter change in R-TLE than in L-TLE patients with respect to controls using voxel-based morphometry.

The average accuracy for T1 and T2 subspaces across models was 74% and 73% respectively. For FA and MD, it was 76% and 74%. Similarly to the classification of L-TLE patients, the best accuracy on individual image spaces was obtained on the MD space. The highest accuracy on the universal feature space was 88.9% and it was obtained using a *correlation-svm-linear* model with 141 features. Intuitively, models with such a high number of features are less useful to the researcher than simpler models. One of such alternatives is presented by the *correlation-pca-svm-linear* model on T1 with K=29 and an accuracy of 81.5%.

Figures 7 and 8 (III) show the regions that were common to all subjects, reported by the aforementioned simpler T1 model. Regions selected by virtue of their *mean intensity* are show in in Figure 7. These regions are the right inferiotemporal cortex, the right entorhinal and parahippocampal white matter regions as well as the right white matter in temporal pole. Neither extra-temporal nor left-temporal regions were chosen.

When looking at regions selected based on *asymmetry*, Figure 8 reveals numerous features most of them corresponding to the temporal lobes, including the inferiotemporal cortex, the parahippocampal white matter and in general the temporal lobe white matter. The difficulty of the classification between R-TLE

and controls explains why the optimal results are obtained by the inclusion extra-temporal features. To validate this assumption we restricted the method to select only temporal lobe features and we found that the classification accuracy reduced from 81.5% to 70.4%.

3.4. Evaluation of feature selection methods

After obtaining the classification results, we performed an analysis of the two feature selection strategies. The two methods were compared measuring their *stability* and *similarity*. To measure the stability of each method we calculated the *average Tanimoto distance* among feature sets [38, 39] in the outer leave-one-out cross-validation loop (training/testing loop). Figure 10 shows the results. As expected, stability increases asymptotically towards 100% as the number of requested features approaches the number of available features. The [2<K<100] range of cardinalities is most interesting as stability varies between 40% and 70% (right column in the figure). This range gives a better estimation of stability as the methods are asked to retrieve at most 100 features out of 936 *for each iteration* of the cross-validation loop. In comparison, in the absence of any heuristic, the likelihood of selecting 100 features out of 936 for the same 36 subjects is approximately 10^{-36}. As shown in the figure, the most stable features ($> 80\%$) were obtained in experiment II (L-TLE) and the least stable features ($< 70\%$) in experiment III (R-TLE). In all three experiments it was observed that classification accuracy was proportional to feature stability.

To assess similarity between methods, in a given training fold, the Tanimoto distance between the ANOVA-generated feature set and the correlation-generated feature set was measured. This was repeated for all the training folds generated by the LOOCV procedure. Then, the average Tanimoto distance was reported. Figure 11 shows how similarity varies according to cardinality. For very small cardinalities (K<20) the two methods (correlation, ANOVA) generate very similar features. In the [20<K<100] range (right column in the figure) the similarity drops to 50%. This is desirable as it shows that the two methods are sufficiently distinct in the cardinality range that contains both the best classification results and the inherent dimensionality of the problem (see the discussion section). Conceptually, the dissimilarity between the two methods can be explained by the redundancy evaluation step in the correlation method. However, as the cardinality approaches the number of available features, the correlation method runs out of features to discard and the similarity between methods increases.

3.5. Relevant features for TLE detection

Feature relevance was evaluated by analyzing how frequently a feature was selected when the classifier was successful. The level of success was defined by setting a minimum accuracy threshold, and the analysis was restricted to those cardinalities where the classifier performance was higher than the threshold. For each cardinality the feature frequency was measured on the external LOOCV (training/testing) analyzing the features obtained from each training fold. Then, a ranking was obtained by averaging frequencies across the

selected cardinalities. The ROI pointed at by the top ranked relevant features are analyzed in the discussion section.

Alternatively, we considered using the coefficient assigned to features by the linear classifiers as a measurement of their relevance. However this approach would not reflect the importance of the features in non-linear classifiers or in classifiers that use PCA. We think that our approach is more comprehensive as it identifies features that are relevant and stable across cardinalities regardless of the type of classification model..

3.6. Effect of the sample size on the reliability of the results

We quantified the effect of the *small sample size* in the performance of the classifier by evaluating the *SVM reliability index (SRI) [40].* Let $\mathbf{w}$ be the vector defining the decision boundary (vector normal to the decision plane), this vector is obtained as the solution of the respective convex optimization problem. Let $\mathbf{w}^*$ be the alternative convex optimization solution after randomly removing some data points. If we have *enough data for training* then we can randomly remove some and what is left will result in $\mathbf{w}^* \approx \mathbf{w}$. If we do not have enough data, the random removal of training data will result in a very different decision boundary and $\mathbf{w}^* \neq \mathbf{w}$. This is quantified by the SRI as follows: R1c

$$SRI(\mathbf{w}^*, \mathbf{w}) = |r(\mathbf{w}^*, \mathbf{w})| \tag{11}$$

which is the absolute value of the Pearson product-moment correlation coefficient between $\mathbf{w}^*$ and $\mathbf{w}$. The SRI was evaluated for each experiment by selecting a training set to estimate $\mathbf{w}$, then the training set was randomly subsampled 10 times to obtain $\mathbf{w}^*$ estimates. Each SRI result was averaged. This process was repeated for all training sets.

It is expected that the decision boundary remains stable when data points that do not weight in its calculation are removed. However, when support vectors are eliminated, this causes the boundary to be redefined and consequently the SRI decreases. Our results show that the decision boundary is fairly sensitive to the training set size (Figure 15). However, in all three experiments we obtained SRI measures above 80% (arbitrarily set). It is relevant to notice that the stability of the boundary is independent from performance: if two classes are sufficiently distinct (large margin), then they can be linearly separable by *many* possible boundaries. From the PCA projections and the experimental results, we believe this is the case for the L-TLE group vs. controls.

4. Discussion

4.1. Analysis of relevant features

Table 4 summarizes the ROIs associated to relevant features, and Figure 14 summarizes this information graphically grouping by lobe, type of feature, and quantitative MRI parameter. In all three experiments

most of these region belong to the temporal lobe (a total of 71 features in the table). Among all, the middle-temporal, superior-temporal, the temporal pole, and hippocampal ROIs are common to all three classification experiments revealing the importance of these regions for the identification of patients. The selection of the hippocampus is to be expected, since it is the presumed focus in mesial temporal sclerosis (MTS), the most common pathology in this group of patients [41]. Relevant features from the neocortex and adjacent white matter could relate either to changes due to seizure propagation, such as gliosis or neuronal loss, or could be related to the presence of epileptogenic lesions in the neocortex, considering that the patient group also included subjects with focal cortical dysplasia (FCD). The inclusion of some extra-temporal regions in the set of relevant features (in the frontal lobe and parietal lobes, as well as the occipital lobe) is also remarkable, however, here the features likely relate to white matter abnormalities related to seizure propagation, and not extra-temporal epileptogenic lesions, since extra-temporal onset was not observed clinically in these patients.

We also analyzed relevant features based on the type of feature. We found that asymmetry features (ks) were chosen more often than mean intensity features from each hemisphere, highlighting the benefit of sensitive examination of intensity distributions between bilateral regions. In patients with unilateral TLE was expected that asymmetry features were highly relevant since seizure onset is restricted to one hemisphere. However, asymmetry features may also be sensitive to compensatory mechanisms occurring in the contralateral hemisphere and thus may not be specific to seizure-related abnormalities. When examining individual image spaces, the classifiers performed optimally in the T1 and MD subspaces, and generally good in FA and T2. Also, classifiers on the universal space obtained accuracies that were better or close to the best individual image subspace. We also see that in experiments I and II, MD features were chosen most, followed by FA. The lateralization ability of diffusion metrics has been shown before [6, 7, 8], including in our previous work, [42], where DTI and relaxometry quantitative imaging parameters were compared in the temporal lobe. This work extends this previous findings showing how these quantitative imaging parameters across the entire brain can be used to classify patient groups.

4.2. Criteria for the selection of the feature set cardinality

The proposed method requires that the feature set cardinality K is set *a priori*. To determine an adequate range for K we plotted K against classification accuracy. For this analysis the best model for each experiment was selected. Due to the heuristic of the feature selection methods described in this paper, relevant features are selected *first*, then, as expected, the accuracy degrades as noisy/non-relevant features are added. Figure 12 shows that above 200 features the average classification approaches the line of random chance. The classifiers present an acceptable behavior with several peaks above the 80% classification rate for $K < 50$,. Given that these plots are taken on the best models we used $K = 50$ as an upper bound for the selection of K.

Independently, a cross-validated L1-penalized logistic regression model was used to estimate K for experiments I, II and III. The number of non-zero coefficients in the regularized regression (regularization factor = 0.1) corresponds to the estimated size of the feature set K. The results are presented in Table 5. In all cases the estimated K is lower than the upper bound previously obtained by simple inspection ($K < 50$). Also, it is important to notice that the estimated K (as shown in Table 5) is close to the number of observations (36 subjects).

4.3. Criteria for the selection of a classification model

Considering classification accuracy, correlation-based models performed slightly better than those based upon ANOVA. This could be explained by the fact that the ANOVA-based method does not eliminate superfluous (collinear) features. Though in theory feature redundancy should not affect classification accuracy, the presence of redundant features in the feature set could hinder the algorithm from finding an optimal set before the desired cardinality is reached.

In terms of the type of SVM, we found that linear classifiers exhibited slightly better performance than non-linear ones in our dataset. However this could be attributable to the size of our sample, which may not be large enough to adequately estimate non-linear SVM parameters. R2

In terms of computation time the ANOVA-based classifiers outperformed the correlation-based classifiers for large cardinalities (Figure 13). This estimation was performed on a machine with 4 CPU cores (Intel Core I7-2600 CPU @ 3.4GHZ) running Ubuntu Linux 12.04 with 16GB of RAM. The ANOVA-based method can evaluate several features simultaneously and it does not incur in the computational cost of the feature redundancy evaluation. Additionally, the PCA transformation and the parameter optimization for non-linear classifier increased significantly the time required to fully classify a dataset.

With these considerations in mind, a good compromise between classification accuracy and computation time can be achieved by using an *anova-pca-linear* model. A linear SVM requires less parameter tuning than a non-linear SVM and classification accuracy did not degrade as rapidly as it did for non-linear classifiers as the cardinality increased. The ANOVA approach is more scalable than the correlation-based method in terms of computational cost when evaluating large feature spaces. In addition, the PCA transformation has the effect of decorrelating the resulting feature space which explains why the classification accuracy is similar to those classifiers using the correlation-based method where redundant features are discarded.

4.4. Comparison with similar studies

Our results were comparable and in some cases better than to those obtained by Focke et al [12]. For example, the classification accuracy for L-TLE described by their method varied between 93% and 95%, while our method obtained perfect identification. In contrast, their accuracy for R-TLE detection was 97% while ours was 88.9%. We believe that the size of the patient group is a key factor to obtain higher accuracies

in the R-TLE group. Their patient group was more than twice the size than ours (38 vs. 17). In addition, we did not perform any Morphometry-based features (such as gray matter and white matter probability maps); these features provided the best performance for their R-TLE group. The purpose of this paper was to explore classification using quantitative imaging (T1 maps, T2 maps, DTI). We plan to investigate the added benefit of adding Morphometry to these features as we believe that this could improve our results. Otherwise, in agreement with these findings, we observed that MD data is more useful for discrimination of TLE patients than FA data, and we also observed poor classification results when looking at T2 data only as they did. Similarly, Keihaninejad et al. [13] stated an accuracy of 86% for the identification of TLE patients where ours is 88.9%, with the caveat that this number refers to MR-negative patients (MRI without clinical findings), therefore more difficult to classify.

4.5. Limitations of this work

It is well known that SVM performance is highly dependent on the quality of the training set. This is a common concern when dealing with biological data characterized by a large number of features and a small number of observations like ours. We have followed several steps to address this limitation and mitigate its effect. R1a

On the one hand, it is clear that the number of features has a direct impact on classification error [43]. Feature sets with cardinalities above the *intrinsic dimension* of the problem can lead to overfitting effects because the classifier can produce decision boundaries that follow the sample points too closely [44]. This effect is evident in our dataset. Figure 12 shows that the generalization ability of the classifier suffers as the cardinality of the feature set increases. A simple method to estimate the ideal cardinality is provided (4.2) and the experiments corroborated optimal results in regions around this estimation.

On the other hand, a small sample size makes necessary the systematic use of cross-validation to avoid overfitting. It is well known that cross-validation reduces the error bias at the expense of increasing the error variance. We believe that we can obtain a better bias-variance trade-off by acquiring more data and reducing the amount of cross-validation. Also, a larger dataset can contribute to SVM parameter optimization, boosting the comparison between linear and non-linear models. R1b

Another limitation in our work is the sensitivity for the detection of R-TLE patients. A larger dataset could potentially improve feature selection leading to better classification rates. Nonetheless, there is evidence suggesting that the R-TLE group is very heterogeneous in terms of MRI abnormalities [45, 46] and the low sensitivity of our results could be associated with that heterogeneity. To validate this, a larger dataset would allow the R-TLE group to be subdivided into specific pathologies (i.e. MTS, non-MTS, FCD) for classification analyses.

5. Conclusion

This paper describes a novel approach for the detection of TLE patients using feature selection and support vector machine methods. The novelty of this work consists of the use of multi-parametric quantitative imaging, the definition of a measurement of regional asymmetry using a non-parametric statistical test and the evaluation and validation of the optimal cardinality of the problem. The main contribution of this paper comprises the evaluation of key factors influencing classification performance, namely the feature selection method, the possibility of a further dimensionality reduction step with the use of PCA, and the type of support vector machine for this type of data. Then, relevant features are identified and their clinical significance is addressed.

Our results demonstrate that the identification of TLE subjects based on quantitative MR images is possible. In particular, DTI derived features seemed to be more effective than T2 features, which in general performed sub-optimally. In all experiments, good accuracies were attainable. However, the identification of R-TLE patients proved to be a difficult problem and in this case, the multi-parametric approach proved to be slightly better (88.9% accuracy) than the classification on individual feature spaces. The subsequent feature analysis confirmed that the key ROIs for patient identification do indeed belong to the temporal lobe. Their relevance in TLE has been indicated by clinical findings and similar research studies. These results reflect the sensitivity of quantitative imaging and the utility of the presented method towards the detection of TLE.

Classification models including PCA transformation after feature selection in general provided better results than the non-PCA models. Among the 8 classification models evaluated, the *anova-pca-linear* model demonstrated the best balance between classification performance and computation time, both key elements in the analysis large datasets. Although SVM behaves coherently in small sample scenarios, a larger patient sample would allow the amount of internal cross-validation and sub-sampling to be reduced. In machine learning terms, this could lead to a better balance between bias and variance.

6. Conflict of interests statement

The authors do not have any conflict of interests to disclose.

7. Acknowledgements

This work is supported by The Canadian Institutes for Health Research , Grant MOP 184807, Canadian Foundation for Innovation, Leading Edge Fund 20994, and the CIHR Post-doctoral research Fellowship 276108 (A. R. Khan).

References

[1] J. F. Tellez-Zenteno, L. Hernandez-Ronquillo, A review of the epidemiology of temporal lobe epilepsy, Epilepsy Research and Treatment 2012 (2012) 5.

[2] R. E. Hogan, R. D. Bucholz, S. Joshi, Hippocampal deformation-based shape analysis in epilepsy and unilateral mesial temporal sclerosis, Epilepsia 44 (6) (2003) 800–806.

[3] N. Bernasconi, D. Kinay, F. Andermann, S. Antel, A. Bernasconi, Analysis of shape and positioning of the hippocampal formation: an MRI study in patients with partial epilepsy and healthy controls, Brain 128 (10) (2005) 2442–2452.

[4] M. Esmaeilzadeh, H. Soltanian-Zadeh, K. Jafari-Khouzani, Mesial temporal lobe epilepsy lateralization using SPHARM-based features of hippocampus and SVM, in: Proc. SPIE, Vol. 8314, 2012, pp. 83144H–83144H–10.

[5] A. Coan, B. Kubota, F. Bergo, B. Campos, F. Cendes, 3T MRI Quantification of Hippocampal Volume and Signal in Mesial Temporal Lobe Epilepsy Improves Detection of Hippocampal Sclerosis, American Journal of Neuroradiology.

[6] H. Li, Z. Xue, M. Dulay, A. Verma, S. Wong, C. Karmonik, R. Grossman, S. Wong, Distinguishing left or right temporal lobe epilepsy from controls using fractional anisotropy asymmetry analysis, in: H. Liao, P. Edwards, X. Pan, Y. Fan, G.-Z. Yang (Eds.), Medical Imaging and Augmented Reality, Vol. 6326 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2010, pp. 219–227.

[7] M. Ahmadi, D. Hagler, C. McDonald, E. Tecoma, V. Iragui, A. Dale, E. Halgren, Side matters: Diffusion tensor imaging tractography in left and right temporal lobe epilepsy, American Journal of Neuroradiology 30 (9) (2009) 1740–1747.

[8] D. W. Gross, Diffusion tensor imaging in temporal lobe epilepsy, Epilepsia 52 (2011) 32–34.

[9] T. M. Ellmore, M. S. Beauchamp, J. I. Breier, J. D. Slater, G. P. Kalamangalam, T. J. O'Neill, M. A. Disano, N. Tandon, Temporal lobe white matter asymmetry and language laterality in epilepsy patients, NeuroImage 49 (3) (2010) 2033 – 2044.

[10] B. E. Boser, I. M. Guyon, V. N. Vapnik, A training algorithm for optimal margin classifiers (1992).

[11] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.

[12] N. K. Focke, M. Yogarajah, M. R. Symms, O. Gruber, W. Paulus, J. S. Duncan, Automated MR image classification in temporal lobe epilepsy, NeuroImage 59 (1) (2012) 356 – 362.

[13] S. Keihaninejad, R. A. Heckemann, I. S. Gousias, J. V. Hajnal, J. S. Duncan, P. Aljabar, D. Rueckert, A. Hammers, Classification and lateralization of temporal lobe epilepsies with and without hippocampal atrophy based on whole-brain automatic MRI segmentation, PLoS ONE 7 (4) (2012) e33096.

[14] K. Chang, H. Jara, Applications of quantitative T1, T2 and proton density to diagnosis, Applied Radiology 34 (1).

[15] S. C. Deoni, B. K. Rutt, T. M. Peters, Rapid combined T1 and T2 mapping using gradient recalled acquisition in the steady state, Magnetic Resonance in Medicine 49 (3) (2003) 515–526.

[16] S. C. L. Deoni, T. M. Peters, B. K. Rutt, High-resolution T1 and T2 mapping of the brain in a clinically acceptable time with despot1 and despot2, Magnetic Resonance in Medicine 53 (1) (2005) 237–241.

[17] S. C. Deoni, High-resolution T1 mapping of the brain at 3T with driven equilibrium single pulse observation of T1 with high-speed incorporation of rf field inhomogeneities (DESPOT1-HIFI), Journal of Magnetic Resonance Imaging 26 (4) (2007) 1106–1111.

[18] S. C. Deoni, Transverse relaxation time (T2) mapping in the brain with off-resonance correction using phase-cycled steady-state free precession imaging, Journal of Magnetic Resonance Imaging 30 (2) (2009) 411–417.

[19] M. Jenkinson, S. Smith, A global optimisation method for robust affine registration of brain images, Medical Image Analysis 5 (2) (2001) 143 – 156.

[20] M. Beg, M. I. Miller, A. Trouve, L. Younes, Computing large deformation metric mappings via geodesic flows of diffeomorphisms, International Journal of Computer Vision 61 (2) (2005) 139–157.

[21] H. Huang, C. Ceritoglu, X. Li, A. Qiu, M. I. Miller, P. C. van Zijl, S. Mori, Correction of B0 susceptibility induced distortion

in diffusion-weighted images using large-deformation diffeomorphic metric mapping, Magnetic Resonance Imaging 26 (9) (2008) 1294 – 1302.

[22] S. C. Deoni, B. K. Rutt, T. M. Peters, Synthetic T1-weighted brain image generation with incorporated coil intensity correction using DESPOT1, Magnetic Resonance Imaging 24 (9) (2006) 1241 – 1248.

[23] R. Desikan, F. Segonne, B. Fischl, B. Quinn, B. Dickerson, D. Blacker, R. Buckner, A. Dale, R. Maguire, B. Hyman, M. Albert, R. Killiany, An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest, NeuroImage 31 (2).

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine Learning in Python , Journal of Machine Learning Research 12 (2011) 2825–2830.

[25] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.

[26] B. Mwangi, T. Tian, J. Soares, A review of feature reduction techniques in neuroimaging, Neuroinformatics (2013) 1–16.

[27] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, 2nd Edition, Springer Series in Statistics, Springer New York Inc., New York, NY, USA, 2001.

[28] M. A. Hall, Correlation-based Feature Selection for Machine Learning, Tech. rep., University of Waikato (1999).

[29] C. Ding, H. Peng, Minimum redundancy feature selection from microarray gene expression data, Journal of Bioinformatics and Computational Biology 03 (02) (2005) 185–205.

[30] N. Sanchez-Marono, A. Alonso-Betanzos, M. Tombilla-Sanromn, Filter methods for feature selection a comparative study, in: H. Yin, P. Tino, E. Corchado, W. Byrne, X. Yao (Eds.), Intelligent Data Engineering and Automated Learning - IDEAL 2007, Vol. 4881 of Lecture Notes in Computer Science, Springer Berlin Heidelberg, 2007, pp. 178–187.

[31] D. D. Cox, R. L. Savoy, Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex, NeuroImage 19 (2) (2003) 261 – 270.

[32] V. Michel, E. Eger, C. Keribin, B. Thirion, Multiclass sparse bayesian regression for fMRI-based prediction, Journal of Biomedical Imaging 2011 (2011) 2:1–2:13.

[33] V. Michel, A. Gramfort, G. Varoquaux, E. Eger, C. Keribin, B. Thirion, A supervised clustering approach for fMRI-based inference of brain states, Pattern Recognition 45 (6) (2012) 2041 – 2049.

[34] J. Mourão-Miranda, A. L. Bokde, C. Born, H. Hampel, M. Stetter, Classifying brain states and determining the discriminating activation patterns: Support vector machine on functional MRI data, NeuroImage 28 (4) (2005) 980 – 995.

[35] J. E. Jackson, A User's Guide to Principal Components, John Wiley & Sons, Inc., 1991.

[36] A. Golugula, G. Lee, A. Madabhushi, Evaluating feature selection strategies for high dimensional, small sample size datasets, in: Engineering in Medicine and Biology Society,EMBC, 2011 Annual International Conference of the IEEE, 2011, pp. 949–952.

[37] G. Lee, C. Rodriguez, A. Madabhushi, Investigating the efficacy of nonlinear dimensionality reduction schemes in classifying gene and protein expression studies, IEEE/ACM Transactions on Computational Biology and Bioinformatics 5 (3) (2008) 368–384.

[38] A. Kalousis, J. Prados, M. Hilario, Stability of feature selection algorithms, in: Data Mining, Fifth IEEE International Conference on, 2005, pp. 8 pp.–.

[39] A. Kalousis, J. Prados, M. Hilario, Stability of feature selection algorithms: a study on high-dimensional spaces, Knowledge and information systems 12 (1) (2007) 95–116.

[40] P. Matykiewicz, J. Pestian, Effect of small sample size on text categorization with support vector machines, in: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing, Association for Computational Linguistics, 2012, pp. 193–201.

[41] F. Al Sufiani, L. C. Ang, Neuropathology of temporal lobe epilepsy, Epilepsy research and treatment 2012.

[42] A. R. Khan, M. Goubran, S. de Ribaupierre, R. R. Hammond, J. G. Burneo, A. G. Parrent, T. M. Peters, Quantitative relaxometry and diffusion MRI for lateralization in MTS and non-mts temporal lobe epilepsy, Epilepsy Research 108 (3) (2014) 506 – 516.

[43] S. J. Raudys, A. K. Jain, Small sample size effects in statistical pattern recognition: Recommendations for practitioners, IEEE Transactions on pattern analysis and machine intelligence 13 (3) (1991) 252–264.

[44] J. Hua, Z. Xiong, J. Lowey, E. Suh, E. R. Dougherty, Optimal number of features as a function of sample size for various classification rules, Bioinformatics 21 (8) (2005) 1509–1515.

[45] N. Kemmotsu, H. M. Girard, B. C. Bernhardt, L. Bonilha, J. J. Lin, E. S. Tecoma, V. J. Iragui, D. J. Hagler, E. Halgren, C. R. McDonald, Mri analysis in temporal lobe epilepsy: cortical thinning and white matter disruptions are related to side of seizure onset, Epilepsia 52 (12) (2011) 2257–2266.

[46] M. Liu, L. Concha, C. Lebel, C. Beaulieu, D. W. Gross, Mesial temporal sclerosis is linked with more widespread white matter changes in temporal lobe epilepsy, NeuroImage: clinical 1 (1) (2012) 99–105.

A. Figures

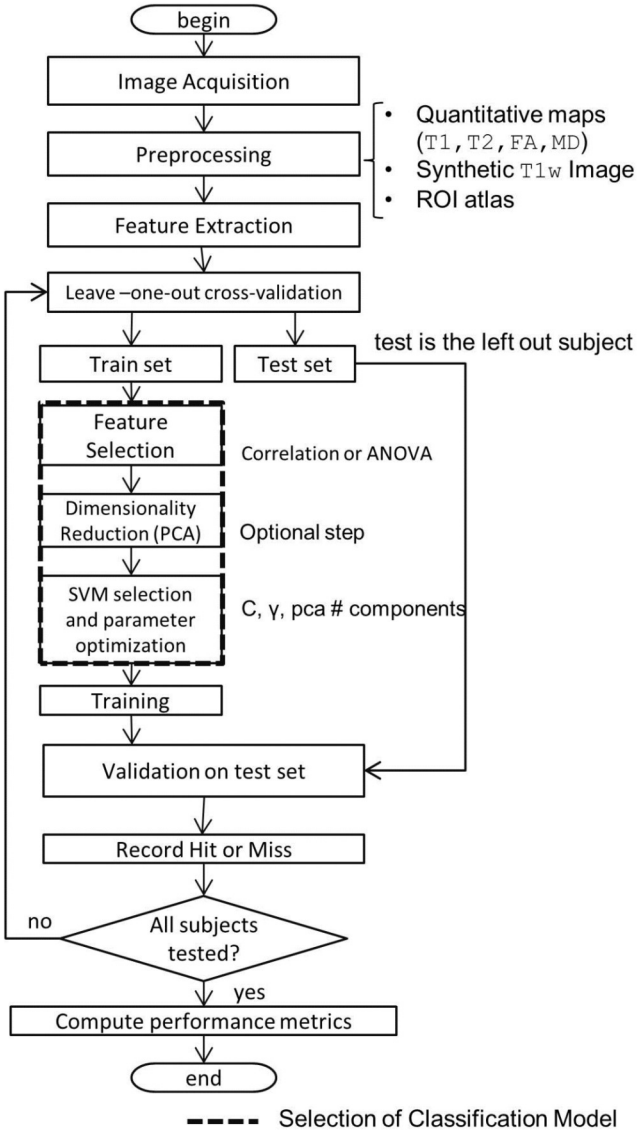


Figure 1: Method. The preprocessing and feature extraction steps are the same for all the experiments and therefore they are executed only once. Each of the experiments described in the current work corresponds to different instances of the leave-one-out cross-validation (LOOCV) loop. The model selection stage determines the feature space, its dimension (PCA-reduced or not), and the parameters that constitute the classification model. 8 models in total are evaluated for each classification experiment. These models are evaluated using performance metrics that are collected at the end of each LOOCV loop.

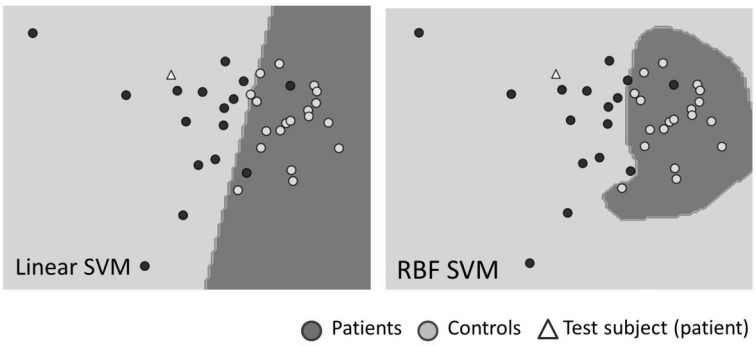


Figure 2: Linear vs. non-linear SVMs. This example shows the decision boundaries of a Linear SVM and a radial basis function(RBF) SVM for the same feature space. The scenario shown is the automatic classification between patients and controls (Experiment I). Here, one of the patients has been selected as the testing set. The classifier is trained using all the other subjects (training set). The patient used as the test set appears to be correctly classified by both the Linear and the RBF SVMs.

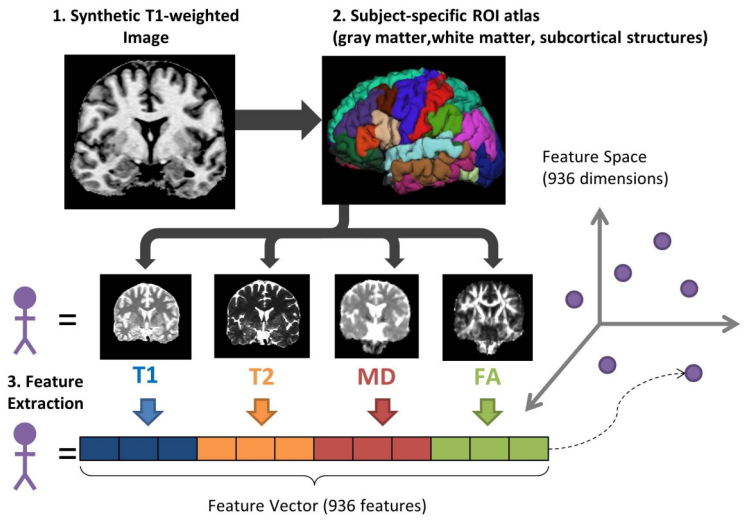


Figure 3: Image Preprocessing and Feature Extraction steps. A subject-specific ROI atlas is employed to extract features from the four different quantitative maps as shown. The resulting feature vector represents each subject in a highly-dimensional feature space. Upon this space is that feature selection algorithms are used to discriminate features that are relevant for classification.

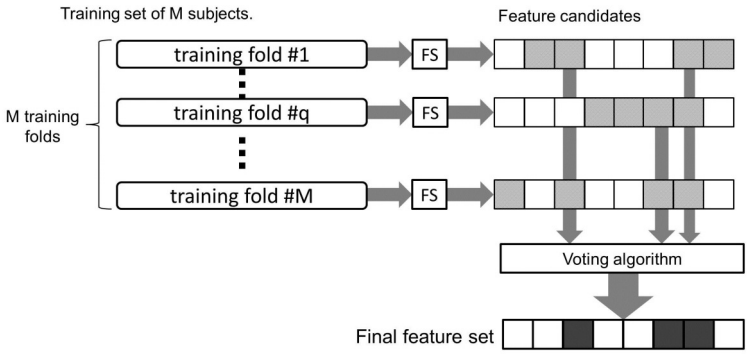


Figure 4: Feature Selection. A training set of size M is subsampled M times by leaving one subject out every time. A feature selection algorithm (FS) is applied on each fold. After feature candidates are obtained, the voting algorithm reviews the agreement among the folds and produces the final feature set that it is used by the SVM.

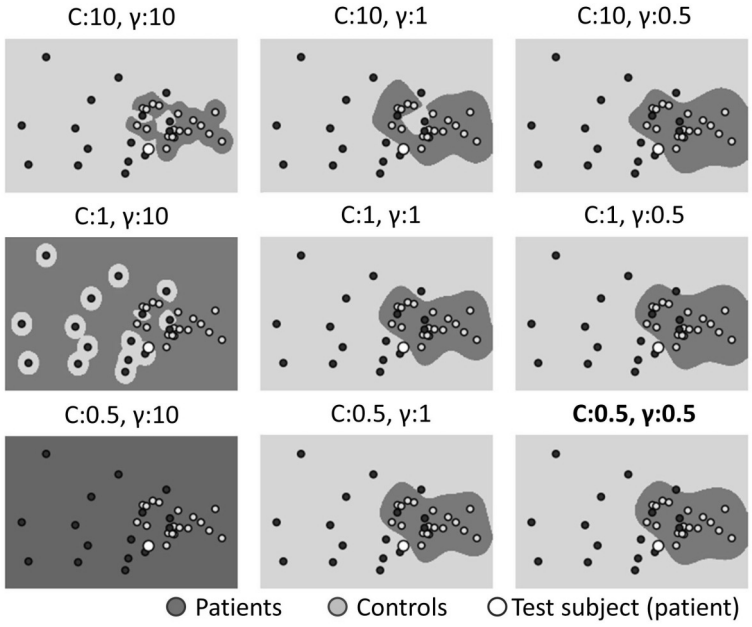


Figure 5: Grid Search for RBF SVM parameter selection. Here, the SVM parameters: C (regularization) and γ (RBF Gaussian kernel width) are explored over the parameter space. An adequate parameter selection allows obtaining a decision boundary that generalizes well and does not overfit. In this example the test subject is correctly included in the patient group while avoiding overfitting which is evident on the three configurations in the first column.

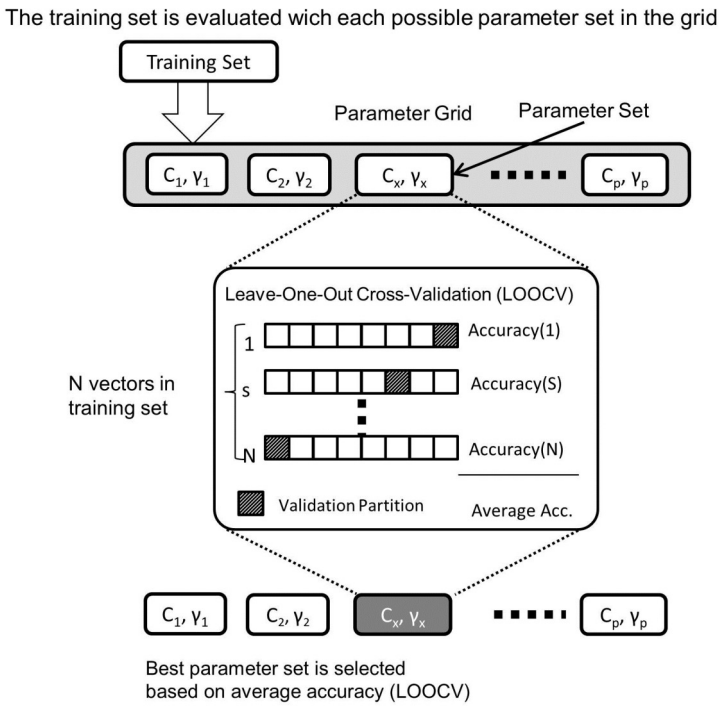


Figure 6: Parameter Selection. A classifier is trained using a training set and one parameter set from the parameter grid. The training set is the same for all classifier instances while the parameter set varies with the goal of evaluating the best parameter configuration for a given type of classifier and training set. A leave-one-out cross-validation procedure is performed on the training set to evaluate the accuracy of the classifier for a given parameter configuration. The best average accuracy determines the set of parameters that will be used. C (regularization) , γ (RBF Gaussian kernel width)

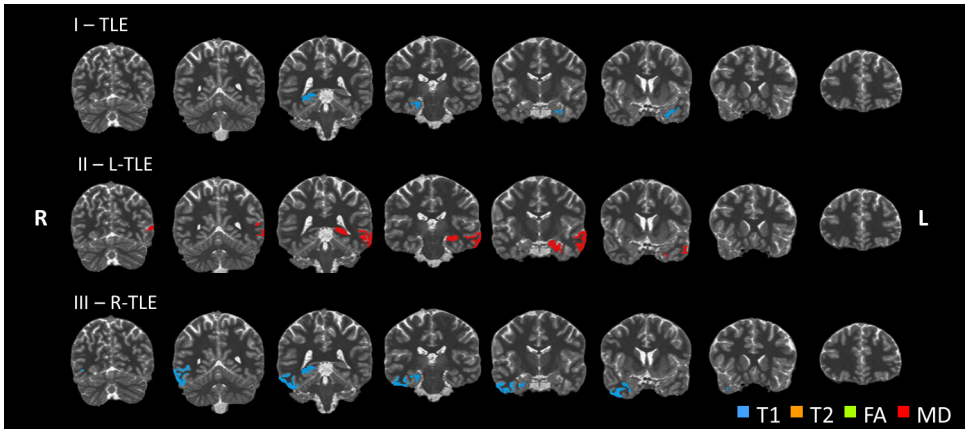


Figure 7: Stable ROI selected based on left/right mean intensity. The best classification cases across subjects are shown. The method used only features in the left temporal cortex to identify L-TLE patients. Similarly, cortical regions in the right temporal lobe were identified as key for classifying R-TLE patients. I) TLE detection: [anova-pca-svm-linear, T1, $K = 10$]. II) L-TLE detection: [correlation-pca-svm-rbf, MD, $K = 7$]. III) R-TLE detection: [correlation-pca-svm-linear,T1, $K = 29$].

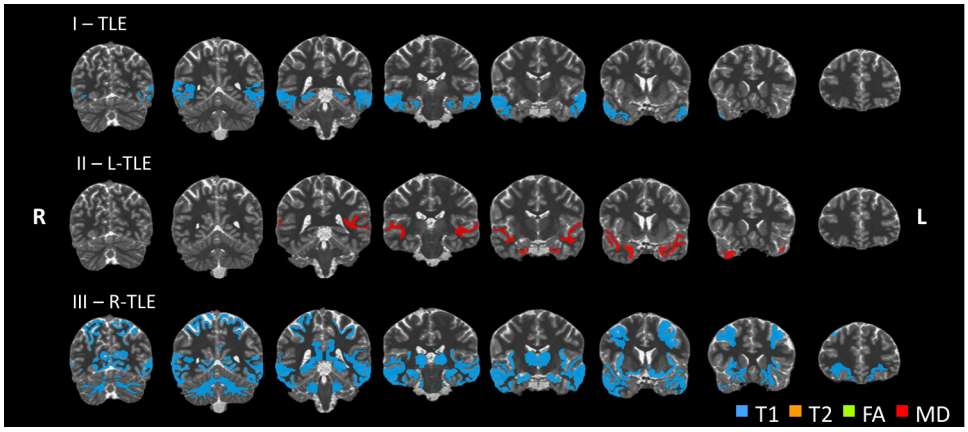


Figure 8: Stable ROI selected based on asymmetry. The best classification cases across subjects are shown. L-TLE patients are successfully classified by looking at the asymmetry in the temporal white matter regions whereas R-TLE patients are harder to classify driving the method to look outside the temporal lobe. I) TLE detection: [anova-pca-svm-linear, T1, $K = 10$]. II) L-TLE detection: [correlation-pca-svm-rbf, MD, $K = 7$]. III) R-TLE detection: [correlation-pca-svm-linear,T1, $K = 29$].

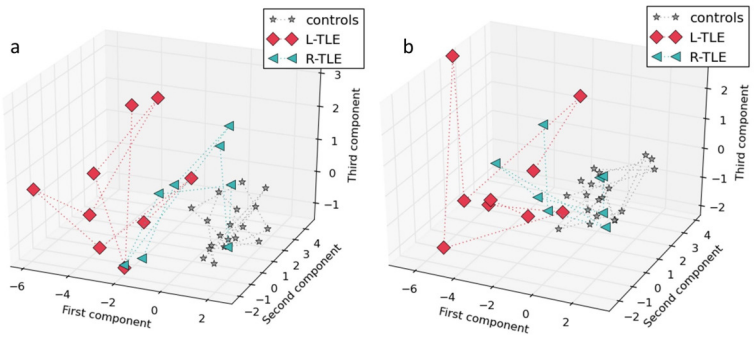


Figure 9: Feature space projection. These tridimensional projections are obtained by performing a PCA transformation from an original feature space with 14 dimensions (K=14) for experiment I. a) using correlation-based feature selection b) using ANOVA-based feature selection. In both projections the control group tends to form a cluster while the patient subgroups are sparse. L-TLE: left TLE, R-TLE: right TLE

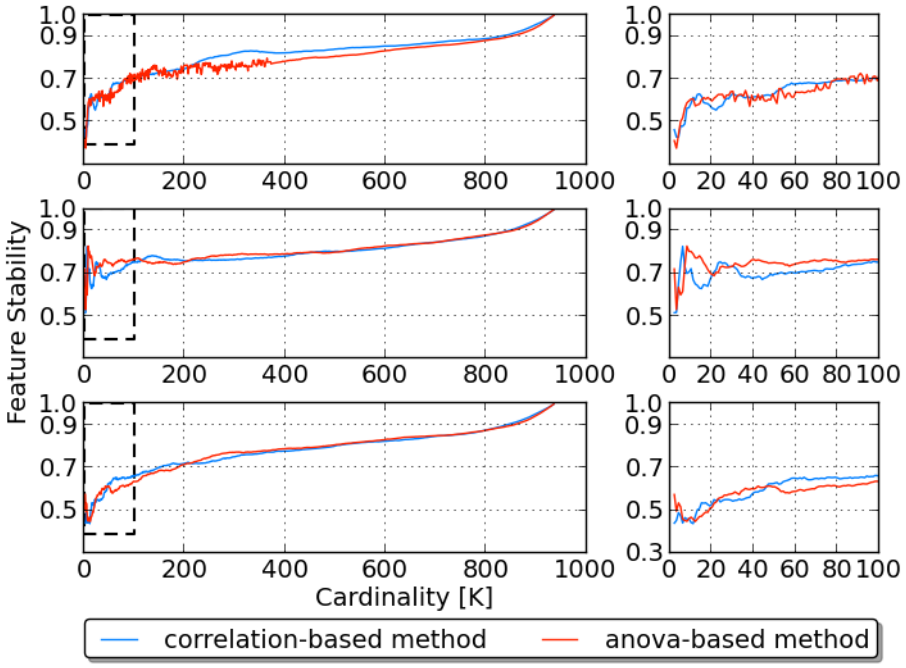


Figure 10: Feature stability. Each training set obtained by leave-one-out cross-validation produces one feature set. The average Tanimoto distance among these sets gives an indication of stability. a) TLE detection b) L-TLE identification c)R-TLE identification. The right column shows a zoomed-in version of the dotted rectangle on the left.

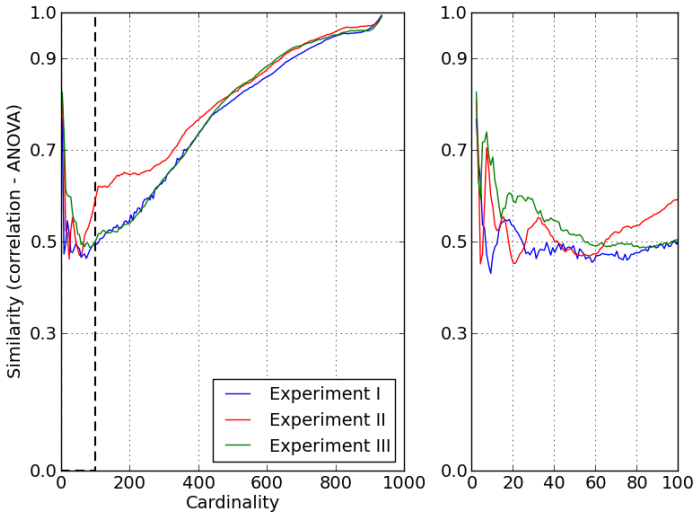


Figure 11: Similarity between the correlation-based and the ANOVA-based feature selection methods. The right column shows the region outlined by the dashed rectangle on the left.

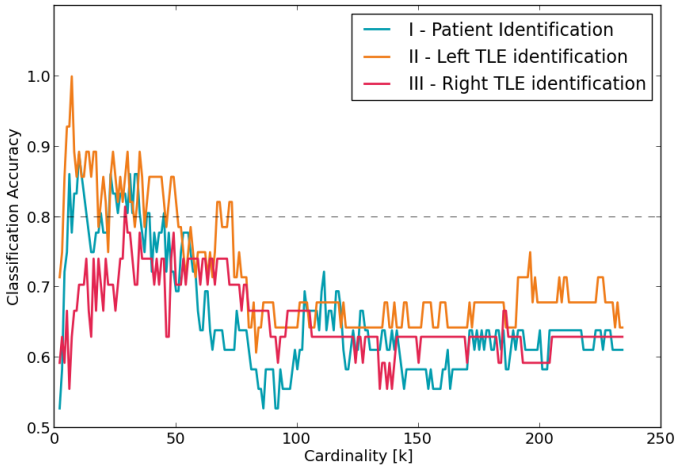
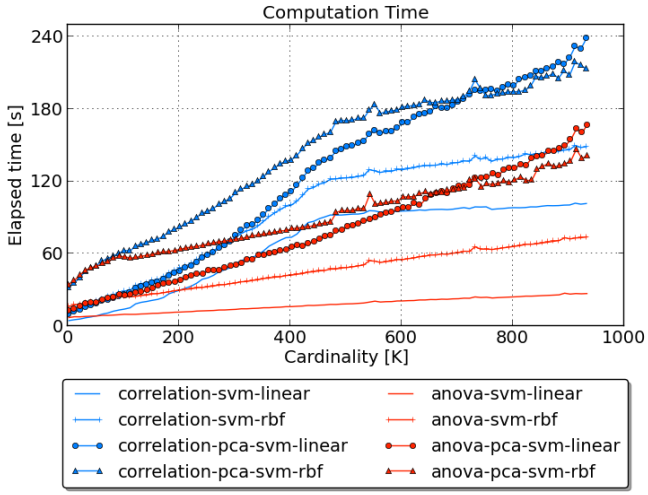


Figure 12: Classification accuracy versus number of features (K parameter). As the requested cardinality increases, the classification accuracy drops. This is expected given that important features are picked first and adding irrelevant features will not improve accuracy but rather will add noise and increase the dimensionality of the feature space beyond the intrinsic dimension of the classification problem.



experment

Figure 13: Performance comparison in terms of computation time for the eight classification models discussed in the paper. The evaluation was made on experiment I, the largest dataset on a machine with 4 CPU cores (Intel Core I7-2600 CPU @ 3.4GHZ) running Ubuntu Linux 12.04 with 16GB of RAM.

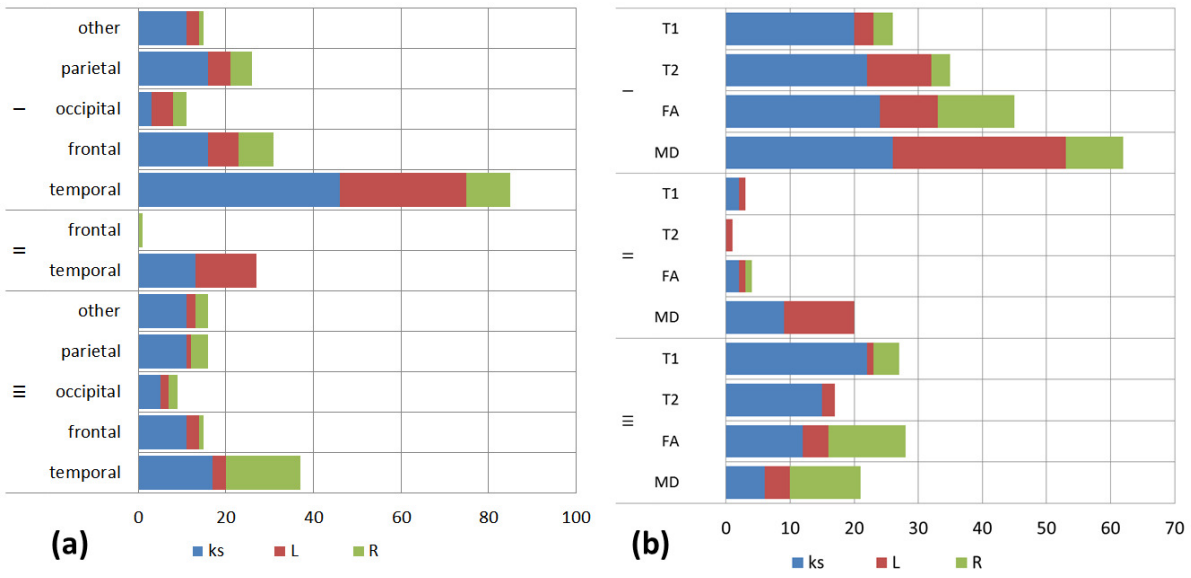


Figure 14: Relevant features for classification by kind. (a) by lobe, (b) by quantitative image. In all three experiments most of the features originated in the temporal lobe. Across all experiments, asymmetry (ks) plays an important role in classification. Mean intensity features in the left hemisphere (L) were preferred over mean intensity features in the right hemisphere (R) to classify L-TLE patients (II). The opposite case was evidenced for the classification of R-TLE patients (III).

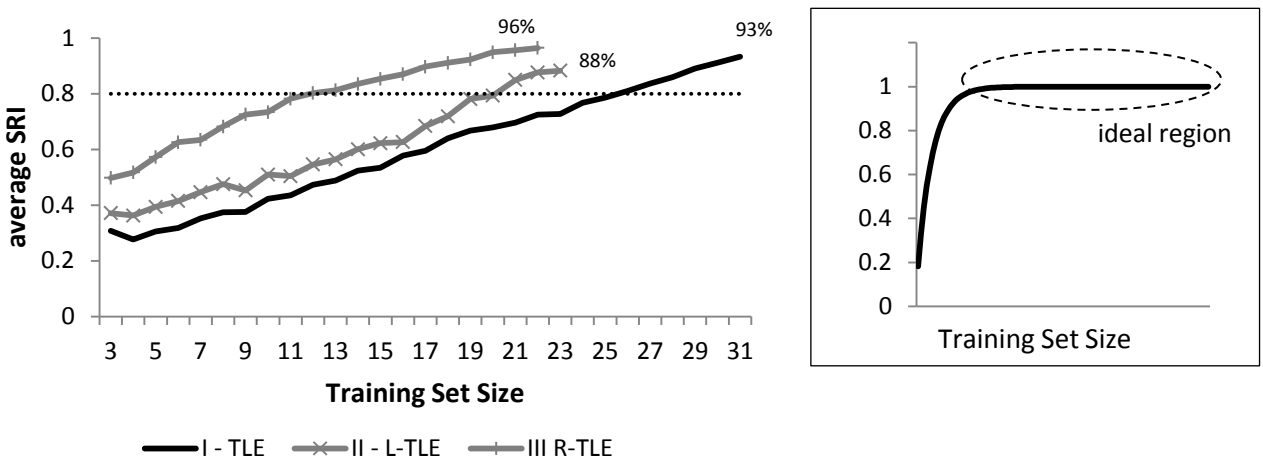


Figure 15: SVM reliability index (SRI). In all three experiments the reliability of the boundary was higher than 80%. A larger dataset would improve the robustness of the boundary with respect to the training set size.

B. Tables

Table 1: Classification results for Experiment I - TLE detection. The best performance for each model is shown. Each column indicates the image from where features were selected. T1: quantitative T1 image, T2: quantitative T2 image, FA: fractional anisotropy image, MD: mean diffusivity image.

Configuration	T1	T2	FA	MD	ALL
correlation-svm-linear					
Accuracy	77.8%	77.8%	69.4%	72.2%	80.6%
Sensitivity	58.8%	76.5%	64.7%	58.8%	64.7%
Specificity	94.7%	78.9%	73.7%	84.2%	94.7%
Cardinality	4	5	22	5	170
correlation-pca-svm-linear					
Accuracy	86.1%	72.2%	72.2%	80.6%	77.8%
Sensitivity	88.2%	76.5%	64.7%	64.7%	64.7%
Specificity	84.2%	68.4%	78.9%	94.7%	89.5%
Cardinality	14	5	47	106	14
correlation-svm-rbf					
Accuracy	77.8%	77.8%	52.8%	77.8%	66.7%
Sensitivity	82.4%	64.7%	0%	82.4%	70.6%
Specificity	73.7%	89.5%	100%	73.7%	63.2%
Cardinality	9	2	14	10	8
correlation-pca-svm-rbf					
Accuracy	80.6%	77.8%	75.0%	80.6%	77.8%
Sensitivity	82.4%	64.7%	76.5%	82.4%	70.6%
Specificity	78.9%	89.5%	73.7%	78.9%	84.2%
Cardinality	6	2	47	59	17
anova-svm-linear					
Accuracy	77.8%	72.2%	72.2%	69.4%	80.6%
Sensitivity	76.5%	64.7%	76.5%	58.8%	64.7%
Specificity	78.9%	78.9%	68.4%	78.9%	94.7%
Cardinality	51	21	37	8	530
anova-pca-svm-linear					
Accuracy	88.9%	66.7%	69.4%	75.0%	75.0%
Sensitivity	82.4%	58.8%	58.8%	58.8%	64.7%
Specificity	94.7%	73.7%	78.9%	89.5%	84.2%
Cardinality	10	12	27	8	13
anova-svm-rbf					
Accuracy	75.0%	77.8%	55.6%	72.2%	63.9%
Sensitivity	88.2%	88.2%	52.9%	58.8%	76.5%
Specificity	63.2%	68.4%	57.9%	84.2%	52.6%
Cardinality	7	7	3	7	12
anova-pca-svm-rbf					
Accuracy	80.6%	72.2%	66.7%	75.0%	77.8%
Sensitivity	82.4%	70.6%	58.8%	58.8%	70.6%
Specificity	78.9%	73.7%	73.7%	89.5%	84.2%
Cardinality	11	7	18	8	16

Table 2: Classification results for Experiment II - Detection of L-TLE patients. The best performance for each model is shown. The best classification spaces were MD and T1. In general, linear models are not outperformed by their non-linear versions. Overall, PCA improves classification accuracy.

Configuration	T1	T2	FA	MD	ALL
correlation-svm-linear					
Accuracy	92.9%	78.6%	85.7%	96.4%	89.3%
Sensitivity	100%	55.6%	66.7%	88.9%	66.7%
Specificity	89.5%	89.5%	94.7%	100%	100%
Cardinality	7	18	5	7	6
correlation-pca-svm-linear					
Accuracy	92.9%	71.4%	85.7%	96.4%	92.9%
Sensitivity	100%	44.4%	66.7%	88.9%	77.8%
Specificity	89.5%	84.2%	94.7%	100%	100%
Cardinality	4	58	47	40	138
correlation-svm-rbf					
Accuracy	85.7%	67.9%	89.3%	92.9%	92.9%
Sensitivity	77.8%	0%	66.7%	77.8%	77.8%
Specificity	89.5%	100%	100%	100%	100%
Cardinality	4	8	5	5	6
correlation-pca-svm-rbf					
Accuracy	96.4%	75.0%	85.7%	100%	92.9%
Sensitivity	100%	22.2%	55.6%	100%	77.8%
Specificity	94.7%	100%	100%	100%	100%
Cardinality	4	96	3	7	23
anova-svm-linear					
Accuracy	85.7%	78.6%	85.7%	92.9%	89.3%
Sensitivity	88.9%	55.6%	66.7%	88.9%	77.8%
Specificity	84.2%	89.5%	94.7%	94.7%	94.7%
Cardinality	5	41	33	10	9
anova-pca-svm-linear					
Accuracy	92.9%	71.4%	85.7%	96.4%	92.9%
Sensitivity	88.9%	44.4%	66.7%	88.9%	77.8%
Specificity	94.7%	84.2%	94.7%	100%	100%
Cardinality	5	66	16	12	10
anova-svm-rbf					
Accuracy	85.7%	67.9%	75.0%	78.6%	78.6%
Sensitivity	66.7%	0%	22.2%	55.6%	55.6%
Specificity	94.7%	100%	100%	89.5%	89.5%
Cardinality	5	8	6	4	4
anova-pca-svm-rbf					
Accuracy	92.9%	71.4%	82.1%	96.4%	92.9%
Sensitivity	77.8%	22.2%	44.4%	88.9%	77.8%
Specificity	100%	94.7%	100%	100%	100%
Cardinality	4	18	7	17	22

Table 3: Classification results for Experiment III - Detection of R-TLE patients. The best performance for each model is shown. The classification of R-TLE patients is a difficult problem in the proposed feature space. This is reflected by the high cardinality of the optimal solutions as well as the average accuracy of 73% across models.

Configuration	T1	T2	FA	MD	ALL
correlation-svm-linear					
Accuracy	77.8%	77.8%	81.5%	74.1%	88.9%
Sensitivity	37.5%	37.5%	62.5%	37.5%	62.5%
Specificity	94.7%	94.7%	89.5%	89.5%	100%
Cardinality	12	143	52	34	141
correlation-pca-svm-linear					
Accuracy	81.5%	77.8%	81.5%	77.8%	70.4%
Sensitivity	62.5%	37.5%	50%	37.5%	37.5%
Specificity	89.5%	94.7%	94.7%	94.7%	84.2%
Cardinality	29	21	13	56	52
correlation-svm-rbf					
Accuracy	66.7%	63%	66.7%	59.3%	59.3%
Sensitivity	25%	25%	25%	12.5%	25%
Specificity	84.2%	78.9%	84.2%	78.9%	73.3%
Cardinality	8	8	7	7	6
correlation-pca-svm-rbf					
Accuracy	74.1%	74.1%	77.8%	77.8%	74.1%
Sensitivity	25.0%	12.5%	50%	25.0%	12.5%
Specificity	94.7%	100%	89.5%	100%	100%
Cardinality	11	53	13	83	179
anova-svm-linear					
Accuracy	77.8%	74.1%	81.5%	74.1%	77.8%
Sensitivity	37.5%	50%	50%	37.5%	25.0%
Specificity	94.7%	84.2%	94.7%	89.5%	100%
Cardinality	56	9	37	110	156
anova-pca-svm-linear					
Accuracy	77.8%	77.8%	81.5%	81.5%	70.4%
Sensitivity	50%	37.5%	50%	50%	25.0%
Specificity	89.5%	94.7%	94.7%	94.7%	89.5%
Cardinality	29	43	122	66	30
anova-svm-rbf					
Accuracy	63%	66.7%	63%	66.7%	55.6%
Sensitivity	25%	12.5%	12.5%	12.5%	25%
Specificity	78.9%	89.5%	84.2%	89.5%	68.4%
Cardinality	7	9	7	6	8
anova-pca-svm-rbf					
Accuracy	74.1%	70.4%	77.8%	77.8%	77.8%
Sensitivity	37.5%	37.5%	25.0%	25.0%	37.5%
Specificity	89.5%	84.2%	100%	100%	94.7%
Cardinality	7	25	78	12	15

Table 4: Summary of ROI identified by relevant features. The values reported are the number of relevant features referring to the respective region. Most of the features indicate the relevance of temporal lobe regions in the classification experiments (71 in total). The enclosed regions are shared by all three classification experiments.

	EXPERIMENT					EXPERIMENT			
Lobe / ROI	**I-TLE**	**II-L-TLE**	**III-RTLE**	**Total**	**Lobe / ROI**	**I-TLE**	**II-L-TLE**	**III-RTLE**	**Total**
temporal	**44**	**5**	**22**	**71**	**frontal**	**18**		**9**	**27**
bankssts	5		2	7	caudal anterior cingulate	1			1
entorhinal	5	1		6	caudal middle-frontal	2		2	4
fusiform	5		3	8	frontal pole	2		1	3
Hippocampus	2	1	2	5	lateral orbito-frontal	1		2	3
inferior-temporal	8		1	9	medial orbito-frontal	6		2	8
middle-temporal	5	1	5	11	pars opercularis	3			3
para-hippocampal			3	3	pars triangularis	1			1
superior-temporal	7	1	2	10	pre-central			1	1
temporal pole	6	1	3	10	superior-frontal	2		1	3
transverse-temporal	1		1	2	**other**	**10**		**7**	**17**
occipital	**4**		**7**	**11**	Accumbens area	1		3	4
cuneus	1			1	Caudate			1	1
lateral-occipital	3		3	6	Cerebellum	4		2	6
lingual			1	1	insula	2			2
pericalcarine			3	3	Putamen	1			1
parietal	**11**		**9**	**20**	Thalamus-Proper	2		1	3
inferior-parietal	1			1					
isthmuscingulate	1		2	3					
post-central	1		1	2					
precuneus	1		1	2					
superior-parietal	5		2	7					
supramarginal	2		3	5					

Table 5: Estimation of K using a L1-penalized logistic regression model (regularization =0.1)

Experiment	estimated K
I - TLE detection	25
II - L-TLE detection	28
III - R-TLE detection	34

***Highlights (for review)**

Highlights

- Measuring regional asymmetry is fundamental for optimal classification results.
- DTI derived measures seem to be more informative than T2 maps for classification of TLE patients.
- Best classification accuracy for left TLE was 100% and for right TLE was 88.9%.
- Most of the discriminative features belong to the temporal lobes.
- The right TLE group is difficult to distinguish from controls. Possible factors are pathology heterogeneity and a limited sample size.

Author Biography

Diego Cantor-Rivera M.Eng, received his Master's degree in Computer Engineering from Universidad de Los Andes, Bogota, Colombia in 2007. He is currently a PhD candidate in Biomedical Engineering at Western University/Robarts Research Institute in London, Ontario, Canada. His current research deals with the analysis of MR images for the detection and characterization of Temporal Lobe Epilepsy. Other interests are neuroimaging analysis, minimally-invasive, image-guided surgery, computer graphics and efficient volumetric rendering applied to medical images.

Ali R. Khan PhD, received his PhD degree in Engineering Science from Simon Fraser University, Burnaby, Canada, in 2011. He is currently a postdoctoral fellow at the Robarts Research Institute, Western University, in the Virtual Augmentation and Simulation for Surgery and Therapy (VASST) laboratory. His current research interests include medical image computing, image registration and segmentation, with applications in neurodegenerative diseases, epilepsy, and image-guided surgery.

Maged Goubran BMSc, received his bachelor's degree in Medical Biophysics from Western University, London, Canada. He's currently a PhD student in the Biomedical Engineering Program at Western. His current research interests include epilepsy, neurological disorders, image registration and high field MR imaging.

Seyed M. Mirsattari MD, PhD, FRCPC, is an associate professor in the departments of Clinical Neurological Sciences, Medical Biophysics, Diagnostic Radiology & Nuclear Medicine, and Psychology at Western University in London, Ontario, Canada. Dr. Mirsattari is interested in using functional neuroimaging modalities to study epileptogenic areas in patients with medically resistant epilepsy that may benefit from respective surgery. He is also interested in basic science research in animal models of epilepsy. As a clinical epileptologist/electroencephalographer, Dr. Mirsattari treats patients with medication resistant epilepsy and conducts clinical research.

Terry M. Peters PhD, FCCPM, FIEEE is a scientist at Robarts Research Institute and a professor in Medical Imaging, Medical Biophysics and Biomedical Engineering at Western University in London, Ontario, Canada. Dr. Peters' laboratory is concerned with the development and validation of tools that allow surgeons to make efficient use of images, produced by sophisticated 3-D imaging systems, during surgical procedures. The objective of minimally-invasive neurosurgery is to resect or lesion the smallest volume of brain tissue, causing the least trauma to the patient while achieving the desired therapeutic result.

LaTeX Source Files

Click here to download LaTeX Source Files: sources.zip